Enhancing Text Analysis via Dimensionality Reduction

David G. Underhill and Luke K. McDowell
Dept. Computer Science, U.S. Naval Academy
572M Holloway Rd, Annapolis, MD 21402
(410) 293-6800
david.g.underhill@gmail.com,lmcdowel@usna.edu

David J. Marchette and Jeffrey L. Solka
Code Q20, NSWCDD
Dahlgren, VA 22448
(540) 653-1910
{david.marchette,jeffrey.solka}@navy.mil

Abstract

Many applications require analyzing vast amounts of textual data, but the size and inherent noise of such data can make processing very challenging. One approach to these issues is to mathematically reduce the data so as to represent each document using only a few dimensions. Techniques for performing such "dimensionality reduction" (DR) have been well-studied for geometric and numerical data, but more rarely applied to text. In this paper, we examine the impact of five DR techniques on the accuracy of two supervised classifiers on three textual sources. This task mirrors important real world problems, such as classifying web pages or scientific articles. In addition, the accuracy serves as a proxy measure for how well each DR technique preserves the inter-document relationships while vastly reducing the size of the data, facilitating more sophisticated analysis. We show that, for a fixed number of dimensions, DR can be very successful at improving accuracy compared to using the original words as features. Surprisingly, we also find that one of the simplest DR techniques, MDS, is among the most effective. This suggests that textual data may often lie upon a linear manifold where the more complex non-linear DR techniques do not have an advantage.

1 Introduction

Individuals, companies, and governments are surrounded by millions of web pages, communications, and other documents that are potentially relevant, yet in danger of being overlooked amongst all the other data. Analyzing such data is complicated by its representation as natural language text, rather than as a structured database record or a numerical measurement. While such documents can be analyzed with sophisticated natural language processing applications, the challenges of fully understanding text and the daunting amount of data often make statistical text mining techniques more attractive. Such techniques typically encode the content of a document as a vector with thousands of dimensions, one for each useful word in the corpus.

One possible approach to simplifying the analysis of such high dimensional data is to apply some form of "dimensionality reduction" (DR). Techniques for DR all seek to take input points and represent them in a much smaller number of dimensions, while retaining important characteristics of the original data. These techniques have been primarily applied to numerical data, particularly data representing geometric shapes. With a few exceptions (see Section 5), they have not been as well studied for textual data.

This paper analyzes how traditional and more recent DR techniques might be profitably applied to textual data. In particular, we evaluate five such techniques: Principal Components Analysis (PCA) [8], Multi-dimensional Scaling (MDS) [11], Isomap [5], Locally Linear Embedding (LLE) [15], and Lafon's Diffusion Maps (LDM) [4]. To measure their effectiveness with text, we perform document classification using both linear and k-nearest-neighbor classifiers. Document classification is a fundamental and useful text analysis task (e.g., for categorizing newly discovered web pages based on previously labeled pages). In addition, our document classification results indicate how well each DR technique has preserved the interesting inter-document relationships. Thus, these results help to predict which DR techniques might be best suited for pre-processing before performing more complex analysis tasks such as Literature-based discovery [14, 17] or sentiment analysis [13].

Our contributions are as follows. First, we demonstrate that, for a fixed number of dimensions, applying DR before classification can significantly improve accuracy. This enables classifiers to be more efficient and reduces the impact of noisy dimensions on the results. Second, we show the surprising result that some of the simplest techniques perform best. In particular, we find that MDS and the closely related Isomap have the most reliable performance, almost always yielding the best performance compared to other DR techniques. In addition, both MDS and Isomap are frequently able to match or approach the performance of

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2007		**00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Enhancing Text Analysis via Dimensionality Reduction	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Naval Aademy,Computer Science Department ,572 M Holloway Road,Annapolis,MD,21402**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
2007 IEEE International Conference on Information Reuse and Integration, August 13-17 2007, Las Vegas, NV

14. ABSTRACT

Many applications require analyzing vast amounts of textual data, but the size and inherent noise of such data can make processing very challenging. One approach to these issues is to mathematically reduce the data so as to represent each document using only a few dimensions. Techniques for performing such "dimensionality reduction" (DR) have been well-studied for geometric and numerical data, but more rarely applied to text. In this paper, we examine the impact of five DR techniques on the accuracy of two supervised classifiers on three textual sources. This task mirrors important real world problems, such as classifying web pages or scientific articles. In addition, the accuracy serves as a proxy measure for how well each DR technique preserves the inter-document relationships while vastly reducing the size of the data, facilitating more sophisticated analysis. We show that, for a fixed number of dimensions, DR can be very successful at improving accuracy compared to using the original words as features. Surprisingly, we also find that one of the simplest DR techniques, MDS, is among the most effective. This suggests that textual data may often lie upon a linear manifold where the more complex non-linear DR techniques do not have an advantage.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**6**	

a naive data representation that uses thousands of dimensions when they use only 10-20 such dimensions. Finally, using specially modified versions of one of our data sets, we demonstrate that the advantages of MDS and Isomap are even more pronounced when classification is more difficult. Overall, our results show that simple techniques can be highly effective at reducing textual data while maintaining its most interesting properties.

The following section explains the DR techniques that we evaluate. Section 3 outlines our experimental method and Section 4 presents our results. Section 5 compares these findings with related work, and Section 6 concludes.

2 Background

Dimensionality reduction can take many different forms. In *feature selection*, DR consists simply of eliminating less significant features from the data set, e.g., by information gain [21]. We focus instead on DR for *feature extraction*, where new features are created based on some (possibly complex) transformation of the input features. DR methods can also be divided into supervised and unsupervised techniques, where the supervised techniques are informed by labels associated with the input instances (or documents). Because we wish to facilitate text mining applications where such labels are often not known (e.g., with the aforementioned sentiment analysis), we focus on unsupervised DR techniques. Finally, a DR technique is *linear* if a linear transformation converts an input data point to the reduced feature space; otherwise, the technique is *non-linear*.

We consider the following unsupervised DR techniques:

1. Principal Components Analysis (Linear) - PCA is a correlation-based technique that chooses a set of representative dimensions called the principal components based on the degree of variation that they capture from the original data [8]. In particular, PCA computes the output points by performing a singular value decomposition (SVD) on the document covariance matrix and then multiplies the resulting eigenvectors with their corresponding eigenvalues.

2. Metric Multidimensional Scaling (Linear) - MDS focuses on preserving distances between pairs of points [20]. The input is a matrix containing pairwise distances between the original points. MDS performs an eigenvalue decomposition on this matrix in a way that embeds the points in a smaller space while maintaining the relative pairwise relationships [11].

3. Isomap (Non-Linear) - Like MDS, Isomap seeks to preserve the pairwise distances between input points [1]. Its matrix, however, is based on the *geodesic* distance, which is computed by connecting points in a computed nearest-neighbors graph. This process can more accurately represent some local patterns [5]. The final result is produced by running MDS on the modified distance matrix.

4. Lafon's Diffusion Maps (Non-Linear) - LDM performs non-linear transformations on an initial inter-point distance matrix in a way that helps accentuate local relationships [4]. In particular, LDM tries to more strongly connect those points that are connected by multiple paths in a nearest-neighbors graph. As with PCA, the final step is a singular value decomposition.

5. Locally Linear Embedding (Non-Linear) - LLE uses geometric intuition and an assumption that high-dimensional data actually resides on some low-dimensional manifold within the large input space [15]. A reduced embedding is found by translating, rotating, and scaling the existing data based on weights that maintain geometric properties present in a graph of nearest neighbors. The transformed data is then reduced via an eigenvalue decomposition.

Table 1. Data Sets

Name	Categories	# Docs	Avg. Doc. Size
Science News	8	1047	8000 chars
Science & Tech.	7	658	2500 chars
Google News	5	3028	4000 chars

3 Methodology

In our experiments, we begin with a collection of documents, encode the documents into a term-document matrix, then perform dimensionality reduction on this large matrix. Using the reduced matrix, we then classify each document into one of several known categories. Finally, we evaluate the results. Below we elaborate on each of these steps.

3.1 Data Sets

Table 1 shows the three datasets that we used for our experiments. The Science News [16] and Google News (drawn from articles on news.google.com) corpuses each contain over 1,000 articles distributed among a number of well-defined categories. S&T [16] is much more difficult to classify because there is much less information per article - some articles only contain 2-3 sentences. Furthermore, S&T has the smallest ratio of articles to categories.

3.2 Document Encoding

To facilitate further processing, our documents must first be encoded. Typically for text mining, a corpus is encoded

as a matrix where each document is described by a row in a matrix. There are a number of possible matrices and ways to compute them, but based on previous work [12, 16] we adopt the following straight-forward scheme. In this method, each column is a feature that corresponds to a single word found in the corpus. Cell (i, j) is then the TF-IDF [3] score for word j in document i, defined as:

$$T_{i,j} = \frac{t/T}{lg(D/d)}$$

where t is the number of times word j appears in document i, T is the total number of words in document i, d is the total number of documents that contain word j, and D is is the total number of documents in the corpus. The resulting matrix is known as a (weighted) Term-Document Matrix (TDM). In our implementation, before computing the TDM we perform word stemming and also discard any word that does not occur at least three times in the corpus.

3.3 Dimensionality Reduction

The encoded TDM has 600-3000 rows (one for each document) and 5000-11000 columns (one for each stemmed word in the corpus). We next use dimensionality reduction to significantly reduce this number of columns (or features).

In our experiments, we use the five DR techniques described in Section 2. We implemented each technique in Java, then validated each against existing Matlab or R code that we obtained from others. Isomap and LLE require a choice of how many closest neighbors to consider when constructing the nearest neighbor graph; experimentally we found that using 10 neighbors yielded good results.

Each technique outputs the most significant dimensions first, enabling us to examine the impact of only using the first M dimensions for classification. In addition, we also compared against two variants that performed *no* DR. First, None-rand randomly selects and uses M features from the unmodified TDM matrix. Second, None-sort also uses M unmodified features, but selects the N features with the highest average TF-IDF score (the more significant words). These two variants highlight the potential classification performance easily obtainable without performing any DR.

3.4 Classifiers

We evaluated with three commonly-used classifiers:

1. **k-Nearest Neighbor Classifier (kNN):** kNN assigns a category to a document based on the class(es) of the k closest instances (nearest neighbors) in the training data as defined by a similarity function. Our implementation computed similarity using cosine distance and picked the most likely category using similarity-weighted voting among the closest neighbors. Experimentally, we found that different values of k produced similar trends; we report results for one setting that worked well for all DR techniques ($k = 9$).

2. **Linear Classifier:** Our second classifier assumes that each feature is normally distributed, and makes a classification decision based on a linear combination of the features. In particular, the likelihood of document i being in category k is computed as:

   $$f(x_i, \hat{c}_k) = \frac{1}{(2\pi)^{\frac{N}{2}} \cdot |\Sigma|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(x_i - \mu_k) \times \Sigma^{-1} \times (x_i - \mu_k)^T}$$

   where N is the number of features, x_i is the row vector representing the document being classified, μ_k is the vector of feature means for category k from the training set, and Σ is the inter-feature covariance matrix. For the linear classifier, the same Σ is used for every category, which results in two categories being separated by a hyperplane in the feature space.

3. **Quadratic Classifier:** The quadratic classifier is like the linear classifier, but uses category-specific covariance matrixes. Consequently, two categories are separated by more complex non-linear functions in the feature space. We found similar trends as with the linear classifier and omit details for lack of space.

We measure classification accuracy, which is the percentage of documents that were assigned to the correct category. All experiments used a leave-one-out procedure for testing.

4 Experimental Results

Table 2 shows the classification accuracy achieved (as a percentage) after encoding a corpus, performing DR, and then applying the kNN classifier (Table 3 gives results for "Linear"). Each row gives the results for one DR technique, or when using None-sort or None-rand (see Section 3.3). Within a corpus, the column labeled "All" shows the results obtained when the classifier is provided with all available features for each document. For None-sort or None-rand, this is every (stemmed) word (5000-11000 depending on the corpus). For the DR techniques, "All" corresponds to using every dimension considered significant enough to be produced by the DR algorithm; this varies from 700 to 3300 dimensions depending on the technique and corpus. In addition, since each DR technique outputs the final dimensions sorted by their perceived significance, we also present results where the classifier uses only the first 5, 10, 20, 50, 100, or 500 dimensions. For instance, reducing Science News with PCA, then classifying using kNN with only the best 20 dimensions yields an accuracy of 81%. Results within 2% of the best for each column are shown in bold.

Before presenting our primary results, we first explain a few trends. First, performance generally increases as

Table 2. Results with the kNN classifier

	Science News							Science & Technology							Google News						
Num. dims.	5	10	20	50	100	500	All	5	10	20	50	100	500	All	5	10	20	50	100	500	All
PCA	**77**	**81**	**81**	**80**	**76**	47	39	33	45	48	52	48	29	11	24	25	26	29	32	34	21
MDS	72	78	78	**78**	**77**	**78**	**78**	**56**	**59**	**60**	**61**	**59**	**64**	**64**	**82**	**86**	**88**	**89**	**89**	**88**	**88**
Isomap	70	76	76	76	**77**	73	73	**57**	**59**	**61**	**60**	**61**	58	58	**83**	**86**	**86**	**87**	**87**	**85**	84
LLE	65	70	75	75	74	28	17	26	26	25	24	24	23	22	78	80	83	83	84	82	72
LDM	65	74	76	75	**75**	17	19	23	26	24	28	34	13	06	22	23	26	57	60	55	51
None-sort	34	41	54	66	67	75	**78**	27	34	41	47	53	59	61	30	35	50	68	73	83	**89**
None-rand	13	14	17	22	36	43	**78**	12	14	15	24	26	39	61	18	19	23	27	38	57	**89**

Table 3. Results with the Linear classifier

	Science News							Science & Technology							Google News						
Num. dims.	5	10	20	50	100	500	All	5	10	20	50	100	500	All	5	10	20	50	100	500	All
PCA	**77**	**82**	**84**	**87**	**93**	**99**	**99**	31	37	45	58	65	95	95	20	22	23	31	34	38	61
MDS	63	74	77	84	87	**98**	**99**	**50**	**58**	**64**	**72**	**79**	**99**	**99**	**78**	**79**	**82**	**85**	**89**	**95**	**99**
Isomap	66	71	77	81	83	**99**	**99**	**51**	**57**	60	67	76	95	95	**78**	**80**	**80**	82	85	**93**	**98**
LLE	49	60	70	76	81	96	**99**	17	18	31	36	41	89	89	54	58	64	70	75	87	**97**
LDM	59	71	76	82	86	**98**	**99**	22	21	25	34	53	90	90	25	26	27	47	53	62	77
None-sort	29	39	48	51	69	**98**	17	23	30	43	54	68	**99**	07	38	46	54	68	76	91	39
None-rand	08	19	16	34	43	91	17	09	12	15	34	43	72	07	28	25	27	39	47	72	39

the number of dimensions increases. When the number of dimensions is very large (500 or "All"), however, performance may decrease, particularly with kNN. We note that kNN performance in this region would likely improve with a weighted similarity function, though the general trends among DR techniques would be the same. Linear is less susceptible to this problem, because it naturally detects when a particular feature is not helpful in discriminating among the classes. Thus, even when using all dimensions output by a particular DR technique (at most about 1000), performance is maintained or improved. However, even Linear can get overwhelmed by many noisy dimensions [18], as occurs with None-sort and None-rand, where "All" corresponds to 5000-11000 dimensions.

Result 1. Compared to other DR techniques, MDS and Isomap yield the most consistent and reliable classification performance. On S&T and Google News, MDS and Isomap dominate. For instance, with 100 dimensions MDS and Isomap improve on other DR techniques by between 11-37% for kNN and 11-38% for Linear. For these data sets, MDS is always within 2% of the best results, and Isomap almost always does just as well.

Science News is somewhat of an exception. For this corpus PCA yielded the best performance, for both kNN and Linear, especially when the number of dimensions is small. For instance, with just 5 dimensions, PCA achieves 77% accuracy with Linear, while the next best is 66% with Isomap. However, MDS and Isomap improve more rapidly as the number of dimensions increases. Hence by 50 dimensions MDS is within 3% of PCA, and Isomap is not far behind.

Furthermore, with kNN, MDS and Isomap do *not* exhibit large drops in performance when using "All" dimensions, unlike PCA, LLE, and LDM. In general, performance with kNN on DR-processed data starts to drop after 100 dimensions, because the newer (less significant) dimensions have more noise. Since our classifier uses an unweighted similarity function, these dimensions have as much impact as the first dimensions, and the noise causes accuracy to decrease. The MDS algorithm, however, naturally recognizes when there is little benefit to be added from additional dimensions, and as a result they take on very small values that have little impact on the similarity calculations. Isomap utilizes MDS as part of its processing and thus inherits this benefit. The net effect is that MDS and Isomap are very reliable performers across the whole range of reduced dimensions.

In contrast, PCA and LDM, the two SVD-based techniques, both perform very poorly on S&T and Google News when the number of dimensions is small. These techniques appear to be more sensitive to the smaller amounts of data per instance in these data sets (see Table 1). Despite this common property, the two data sets yield significantly different performance. On Google News, peak performance for both kNN and Linear reaches over 80% with just 10 dimensions. This can be attributed to the data set having just five well-separated categories evenly split among over 3,000 articles. Science and Technology, on the other hand, has seven categories and is unevenly distributed over just 658 articles. As a result, S&T needs 500 dimensions to get over 80% accuracy with Linear and peaks at 64% for kNN.

Result 2. Compared to not performing DR, applying DR before classification can greatly improve the accuracy obtainable with a given number of dimensions. Using MDS or Isomap almost always improved accuracy compared to None-sort or None-rand with the same number of dimensions. For instance, with Linear on Science News with 100 dimensions, using the raw words (None-rand) yielded an accuracy of just 43%. Choosing 100 good words with None-sort improved accuracy, but only to 69%. MDS or Isomap, however, reached 83-87%. For PCA, LLE, and LDM, this improvement held true for Science News, but usually not for the more challenging S&T or Google News.

Result 3. The best DR techniques can enable very strong accuracy using a small number of dimensions. With kNN, MDS with just 10-20 dimensions achieves nearly the same accuracy obtained by any DR or no-DR variant with any number of dimensions. For instance, MDS with 20 dimensions on Google News has an accuracy of 88%, compared to the best found with kNN of 89%. Isomap behaves similarly.

Linear requires many more dimensions (500 or more) to reach its peak performance, though that level is always higher than with kNN. However, MDS and Isomap can still achieve very solid performance with fewer dimensions. For instance, with 100 dimensions MDS reaches 87% for Science News, 79% for S&T, and 89% for Google News.

Result 4. MDS and Isomap's advantages may be even more pronounced on more difficult data sets. To further investigate the situations where MDS and Isomap perform best with regards to classification, we created three variants of the Science News data set. SN2 is a subset of Science News that contains only Medicine and Astronomy articles - two well-separated categories. SN4-Sep contains four fairly well separated categories: Astronomy, Medicine, Earth Sciences, and Life Sciences. Finally, SN4-OL contains four categories with collectively more topic overlap: Medicine, Life Sciences, Behavior, and Anthropology.

In general, we found consistent results as before. Table 4 summarizes the results obtained when using 50 dimensions. SN2 is a much easier task, and almost every technique gets near perfect accuracy. SN4-Sep is somewhat harder, but still yields better results than with the original Science News, and every DR technique yields an accuracy above 80%. With the more difficult SN4-OL, however, PCA drops from 86-89% to just 31-43%. LDM suffers similar losses. MDS and Isomap, however, decline by only 1-4%.

So while every DR technique did well on the easier SN2 and SN4-Sep corpuses, only MDS and Isomap maintained accuracy over 80% on SN4-OL (though LLE was close). Thus, when combined with our results from S&T and Google News, we observe that MDS and Isomap perform at or near the best observed levels, but particularly dominate when the the analysis is complicated by shorter documents or less well-separated categories.

Table 4. Sci. News Variations (# dims. = 50)

	kNN			Linear		
Corpus	SN2	SN4-Sep	SN4-OL	SN2	SN4-Sep	SN4-OL
PCA	98	86	43	99	89	31
MDS	99	85	81	99	89	89
Isomap	99	82	83	99	88	86
LLE	99	84	78	99	87	84
LDM	98	82	44	99	89	50
None-sort	99	81	69	98	87	76
None-rand	85	68	44	83	74	54

5 Related Work and Discussion

As previously discussed, DR has frequently been applied to data sources such as images [1, 5, 15] and biological data [7]. However, the uses of DR for text, particularly the sorts of unsupervised DR methods that we explore in this paper, have not been nearly as extensive. For instance, van der Maaten et al. [19] present a thorough analysis of classification performance using twelve DR techniques on seven data sets, but only one data set is textual. In addition, they do not examine the impact of different numbers of dimensions for the text corpus, do not consider perturbing the corpus, and do not evaluate MDS. Nonetheless, it is useful to make some limited comparisons between our results. We both find that more complex techniques such as LLE and LDM do *not* usually improve over simpler techniques like PCA. However, PCA performs very well for their data sets, particularly on the one text corpus (where it is the best unsupervised technique). Our results suggest that PCA may indeed perform very well on some textual data (e.g., Science News), but poorly on more challenging data sets such as S&T.

Kim et al. [10] describe alternative DR techniques that can reduce the number of features needed to perform text classification. These techniques, however, require that the words in the data set already be clustered together into logical groups. Karypis and Han [9] explored how to expand concept indexing, an alternative DR technique, to exploit known classes of documents. For this one supervised technique, they demonstrate a small improvement in classification accuracy when using the reduced feature space. Bingham and Mannila [2] compare a PCA-like DR technique against a computationally efficient approach based on random projection, and find that the simpler approach is able to achieve comparable performance. However, for text they evaluate only one DR technique, use only one corpus, and evaluate distortion error rather than classification accuracy.

In contrast to the above work, we evaluate classification accuracy using five DR techniques on three distinct corpuses. In addition, we do not require that any class labels or word cluster information be provided to the DR algorithms.

6 Conclusions

This paper has examined how pre-processing with dimensionality reduction could improve text analysis, using classification accuracy to measure performance. Of the five DR techniques that we considered, all were able to achieve substantial improvements compared to not performing DR, under some conditions. However, MDS and the closely related Isomap proved to be the best overall performers. These two techniques consistently out-performed the no-DR variants that we considered, and were able to achieve strong accuracy using just a small number of dimensions.

MDS's ability to out-perform more recent and more complex techniques such as LDM was surprising. This suggests that, at least for the corpuses we considered, this textual data lies on a linear manifold for which the more complex non-linear techniques do not hold an advantage. In addition, we found that LDM, as well as older techniques such as PCA and LLE, had particular trouble with corpuses where the documents were smaller or belonged to less well separated categories. Both MDS and Isomap, however, maintained good performance under these situations. Future work should verify these findings with additional data sets and other classifiers. Additionally, it would be worthwhile to more precisely characterize the corpuses for which MDS and Isomap hold an advantage.

The ability to reduce input documents while preserving significant inter-document relationships is important for several reasons. First, having fewer dimensions enables a classifier to function much more efficiently, improving both training and testing time, and also eliminates many noisy dimensions that could diminish accuracy. Second, informative low-dimension representations can be stored and transmitted more efficiently. Third, low-dimension representations enable high-order processing to be more efficient. For instance, discovering previously unknown connections between documents can be accomplished more quickly on reduced data [14]. Furthermore, effective pre-processing with DR may even improve the quality of such analysis tasks by eliminating noise and by identifying distinct terms with related meaning, as with latent semantic analysis [6]. We intend to further explore such possibilities in future work.

Acknowledgments. Thanks to the USNA Trident Scholar and ONR In-house Laboratory Independent Research Programs for supporting this work, and to Eric Hardisty and the anonymous reviewers for their helpful comments.

References

[1] M. Balasubramanian and E. L. Schwartz. The Isomap algorithm and topological stability. *Science*, 295:7a, Jan. 2002.

[2] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *ACM Special Interest Group on Management of Data*. ACM Press, 2001.

[3] K. W. Church and W. A. Gale. Inverse document frequencies (idf): A measure of deviations from Poisson. In *Annual ACM Conference on Research and Development in Information Retrieval*. ACM Press, 1995.

[4] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS*, 102(21):7426–7431, May 2005.

[5] T. Friedrich. Nonlinear dimensionality reduction - Locally Linear Embedding versus Isomap. Technical report, The University of Sheffield - Machine Learning Group, Sheffield S1 4DP, U.K., December 2004.

[6] M. D. Gordon and S. Dumais. Using latent semantic indexing for Literature Based Discovery. *Journal of the American Society for Information Science*, 49(8):674–685, 1998.

[7] B. W. Higgs, J. Weller, and J. L. Solka. Spectral embedding finds meaningful (relevant) structure in image and microarray data. *BMC Bioinformatics*, 7:74, February 2006.

[8] I. Jolliffe. Principal component analysis. Technical report, Springer, October 2002.

[9] G. Karypis and E.-H. Han. Fast dimensionality reduction algorithm with applications to document retrieval & categorization. In *International Conference on Information and Knowledge Management*, 2000.

[10] H. Kim, P. Howland, and H. Park. Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 6:37–53, b 2005.

[11] J. B. Kruskal and M. Wish. *Multidimensional Scaling*, chapter 1, 3, 5, pages 7–19, 48, 73. Sage Publications Inc., 1978.

[12] A. R. Martinez and E. J. Wegman. A text stream transformation for semantic-based clustering. *Computing Science and Statistics*, 34:184–203, 2002.

[13] T. Nasukawa and J. Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Int. Conference on Knowledge Capture*, Saimbel Island, FL, Oct 2003.

[14] C. E. Priebe, D. J. "Marchette", and al. Iterative denoising for cross-corpus discovery. *COMPSTAT 2004 Symposium*, 2004.

[15] L. K. Saul and S. T. Roweis. An introduction to locally linear embedding. Technical report, AT&T Labs Research and Gatsby Computational Neuroscience Unit, UCL, 2001.

[16] D. J. Solka, A. C. Bryant, and E. J. Wegman. Text data mining with minimal spanning trees. *Handbook of Statistics on Data Mining and Visualization*, 24, 2005.

[17] D. R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18, 1986.

[18] G. Trunk. A problem of dimensionality: A simple example. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 1(3):306–307, July 1979.

[19] L. van der Maaten, E. Postma, and H. van den Herik. Dimensionality reduction: A comparative review. Draft Version., 2007.

[20] F. Wickelmaier. An introduction to MDS. Technical report, Aalborg University (Denmark), May 2003.

[21] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *International Conference on Machine Learning*, 1997.